第9講　データの活用

PART1　データの収集と整理

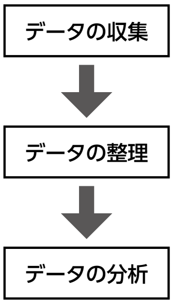
目標
- データ分析の流れについて理解する。
- データの種類や収集方法について，適切に考えられるようになる。

Chapter1　データの収集と整理

データ分析の流れ

　収集した大量のデータは，〔①　　　　〕・〔②　　　　〕することによって未来を予測したり，新たなサービスを生み出したりすることに役立つ。

覚えておこう
データ分析の流れを図解すると以下の通りである。



データの種類

　データを収集するときは，データの種類に注意する必要がある。
　データは大きく〔③　　　　　　　〕と〔④　　　　　　　〕に分類される。そこから更に，データの値を決めるときの基準である〔⑤　　　　〕の違いによって分けられる。

・量的データ

　量的データは，数量的な意味があり計算が可能なデータである。

〔⑥　　　　　〕	〔⑦　　　　　〕
数値の間隔（差）が数量として意味を持つ。 【間隔尺度の例】 ・気温 「昨日の気温は 20 度，今日は 23 度なので 3 度上昇した」というように，間隔（差）の計算ができる。 ・西暦 「2000 年の 23 年後は 2023 年である」というように，間隔（差）の計算ができる。	数値の間隔（差）だけでなく，数値の比にも数量としての意味を持つ。 【比例尺度の例】 ・距離 「10 ｋｍは 2 ｋｍの 5 倍の距離がある」というように，比率に意味を持つ。 ・値段 「商品 A は 200 円，商品 B は 100 円なので 2 倍の差がある」というように，比率に意味を持つ。

知っておこう
比例尺度は，比率尺度とよばれることもある。

重要
間隔尺度の気温 0 度は「気温がない」ということではないが，比例尺度の距離 0 km は「距離がない」という絶対的な意味を持つ。間隔尺度の 0 は相対的であるが，比例尺度の 0 は絶対的であるという違いがある。

・質的データ

質的データは，数値的な意味がなく分類として意味を持つデータである。

〔⑧　　　　　〕	〔⑨　　　　　〕
異なる分類として区別することだけに意味を持つ。 【名義尺度の例】 ・血液型 A 型・B 型・O 型・AB 型という異なる分類として区別することに意味を持つ。 ・都道府県名(都道府県コード) 1 北海道，13 東京都，26 京都府，47 沖縄県など区別することに意味を持つ。	分類の順序に意味を持つ。 間隔は一定ではない。 【順序尺度の例】 ・通知票 5 段階の場合, 1 (最も悪い) ～ 5（最も良い）という順序に意味を持つ。 ・徒競走の順位 1 位, 2 位…という順序（順位）に意味を持つ。

知っておこう

名義尺度は，名目尺度とよばれることもある。

データの収集

・オープンデータ

〔⑩　　　　　　　　〕：国や地方公共団体および事業者などが保有するデータで，次のような項目に該当する形で公開されているもの。

(1) 営利目的，非営利目的を問わず，〔⑪　　　　　　〕なルールが適用されている。

(2)〔⑫　　　　　　〕に適している（コンピュータで容易に処理できるデータ形式)。

(3)〔⑬　　　　〕で利用できる。

例えば，デジタル庁が設置，運営する DATA GO.JP では，統計情報などさまざまなオープンデータが提供されている。

覚えておこう

データ収集・分析の際には，バイアスにも注意をする必要がある。バイアスとは，差別，偏見，誤った認識などにより収集したデータに偏りが生じることである。バイアスが含まれているデータでは，正しい分析結果を得ることは難しい。

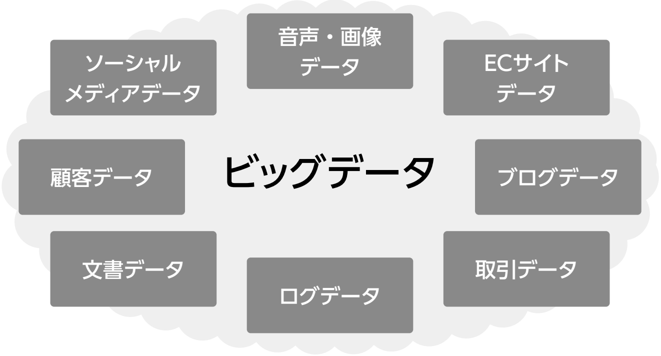
※バイアスとは偏りである。偏りを取り除くことによって正しい分析結果を得ることはできる。しかし，それは困難である。

【引用】DATA GO.JP
データカタログサイト

- **ビッグデータ**

　さまざまな形をした，さまざまな性格を持った，さまざまな種類の膨大なデータのことを〔⑭　　　　　　　　〕という。その特徴は，データの〔⑮　　　〕(Volume)，データの〔⑯　　　〕(Variety)，データの〔⑰　　　　　　　　〕(Velocity)の頭文字から「3V」と表現されることもある。

知っておこう
データを用いて，社会に有益な知見などを科学的に引き出そうとするアプローチのことをデータサイエンスという。



データの整理

- **データクレンジングと前処理**

　収集したデータは，分析に利用できるデータにするために整理する必要がある。例えば，欠損値や外れ値が存在しないか，必要なデータが全てそろっているかなどの確認が必要である。このような，破損したデータ，不正確なデータ，無関係なデータを特定して，必要に応じて修正や削除を行う手法のことを〔⑱　　　　　　　　〕という。また，分析前に，分析ツールなどに合わせて必要な加工をすることもある。

知っておこう
データクレンジングは，データクリーニングと呼ぶこともある。

日付	気温（度）
9月1日	31.2
9月2日	29.5
9月3日	
9月4日	28.2
9月5日	30.1
9月6日	41.5
9月7日	29.2

欠損値
機器の故障，入力忘れ等で欠けている値のこと

外れ値（異常値）
データの中で，他のものと比べて大きく外れた値のこと

重要
外れ値の明確な定義はないが，平均値から標準偏差の 2 倍または 3 倍以上離れている数値を外れ値と見なすことがある。
また，箱ひげ図から外れ値を定義することもある。この場合，第 1 四分位数から第 3 四分位数の距離（箱の長さ）の 1.5 倍以上箱から離れた値を外れ値とすることが多い。

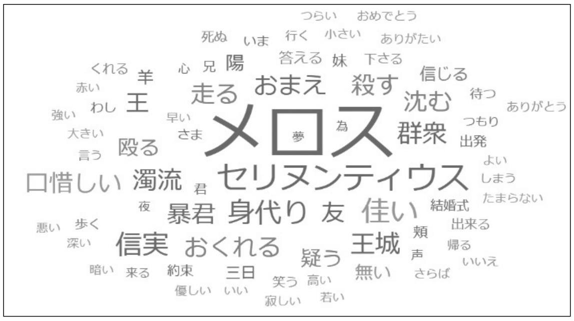
データの分析

・データマイニングとテキストマイニング

〔⑲　　　　　　　　　　　　〕：大量のデータから有用な情報を抽出し，自動的に傾向やパターンを発見する分析手法。

データマイニングのうち，特にテキストデータを対象としたものを〔⑳　　　　　　　　　　　　〕という。

例.太宰治『走れメロス』のテキストマイニング



名詞	スコア	出現頻度	動詞	スコア	出現頻度
メロス	542.18	76	走る	6.43	24
王	12.91	18	くれる	0.53	21
セリヌンティウス	110.84	15	殺す	5.15	17
おまえ	13.13	15	言う	0.14	13
友	12.62	15	行く	0.13	13
君	1.19	13	出来る	0.38	12
妹	3.03	12	殴る	3.52	11
いま	1.30	12	信じる	1.77	11
声	0.37	9	死ぬ	0.35	10
陽	8.19	8	来る	0.16	10
わし	1.57	8	待つ	0.37	9
群衆	31.76	7	帰る	0.21	9
心	0.33	7	沈む	6.65	8
身代り	61.38	6	下さる	0.49	8
濁流	24.52	6	笑う	0.32	8

【引用】UserLocal AIテキストマイニング（ https://textmining.userlocal.jp/ ）

まとめ

・データは大きく，**量的データ**と**質的データ**に分類され，そこから更に，データの値を決める際の基準である**尺度**の違いによって分けられる。

・**データクレンジング**は，破損したデータ，不正確なデータ，無関係なデータを特定して，必要に応じて修正や削除を行う手法のことである。

・**データマイニング**とは，データから有用な情報を抽出し，自動的に傾向やパターンを発見する分析手法のことである。

Chapter2　実践問題

問 1　地域の気候のデータ分析を行うとする。次の①～③をデータ分析の手順としてふさわしい順番に並べ替えなさい。

①　必要なデータがそろっているか，欠損値などが無いかを確認し，データクレンジングを行った。

②　表計算ソフトウェアを使ってデータの分析を行った。

③　気象庁のサイトから，必要な地域の気象データをダウンロードした。

問 2　データの尺度について，ア～エのデータは，それぞれどの尺度に該当するか。①～④から選びなさい。

ア　服のサイズ（S サイズ，M サイズなど）

イ　部屋の気温

ウ　出身地

エ　歩いた距離

①　間隔尺度　　②　比例尺度

③　名義尺度　　④　順序尺度

問３　次のア～エの説明について，正しければ○，誤っていれば×を答えなさい。

ア　多数の患者のデータについて，データマイニングツールを用いて分析することにより，症状による病気の推測と有益な治療法の提案ができる可能性が高まる。

イ　ビッグデータとは，さまざまな形をした，さまざまな性格を持った，さまざまな種類の膨大なデータのことで，これらを分析することで，新たな知見を得られる可能性が高まる。

ウ　オープンデータを利用するとき，一部のデータについては利用料を支払う必要がある。

エ　データ分析を行うときに，詳細な分析の前に，欠損値・外れ値の有無などを確認し，必要に応じてデータクレンジングを行ったり，分析ツールに合わせた必要な加工を行ったりする。

問４　次のア～エのテキストマイニングの説明から，誤っているものを１つ選びなさい。

ア　SNS への投稿の膨大なデータを分析し，消費者ニーズやキーワードを発見した。

イ　迷惑メールのデータを分析して，迷惑メールフィルタのルールに活用した。

ウ　顧客との対話データを分析して，問い合わせの傾向を把握し，チャットボットの回答パターンに取り入れることで，オペレータの負荷軽減につながった。

エ　講演会の内容を録音したものを自動で文字起こしして，テキストデータとして保存した。

知っておこう

チャットボットとは，チャット（会話）をロボット化することで，顧客とのコミュニケーションを自動化できるツールのこと。

第9講　PART1　確認問題

問１　データは，大きく分けて量的データと質的データの２種類に分類できる。量的データにはａ間隔尺度とｂ比例尺度，質的データにはｃ名義尺度とｄ順序尺度の，それぞれ２つの尺度がある。

(1)　下線部ａの例として最も適切なものを，次の①～④のうちから一つ選びなさい。

①　通知票の５段階評価　　②　西暦
③　身長　　④　血液型

(2)　下線部ｂの例として最も適切なものを，次の①～④のうちから一つ選びなさい。

①　通知票の５段階評価　　②　西暦
③　身長　　④　血液型

(3)　下線部ｃの例として最も適切なものを，次の①～④のうちから一つ選びなさい。

①　通知票の５段階評価　　②　西暦
③　身長　　④　血液型

(4)　下線部ｄの例として最も適切なものを，次の①～④のうちから一つ選びなさい。

①　通知票の５段階評価　　②　西暦
③　身長　　④　血液型

問２　オープンデータの特徴として最も適切なものを，次の①～④のうちから選びなさい。

① 営利，非営利目的を問わず，二次利用が可能
② コンピュータで処理しづらいデータ形式になっている
③ データの利用は有償である
④ 必ず利用許諾が必要なデータである

問３　ビッグデータの説明として最も適切なものを，次の①～④のうちから選びなさい。

① EC サイトに関する膨大なデータのことである
② 二次利用が可能なデータのことである
③ どのデータも，データの形式はすべて同じである
④ 分析することでさまざまな意思決定に役立てることができる

問４　収集したデータを分析に活用できるようにするために，事前に行う処理のことを（　ア　）という。その後，（　イ　）という分析手法を用いて分析し，大量の情報から有用な情報を抽出して，傾向やパターンなどを発見することができる。

(1)　空欄（ア）に入る語句として最も適切なものを，次の①～④のうちから一つ選びなさい。

① データクレンジング　② データマイニング
③ バイアス　④ フィルター

(2)　空欄（イ）に入る語句として最も適切なものを，次の①～④のうちから一つ選びなさい。

① データクレンジング　② データマイニング
③ バイアス　④ フィルター

第９講　データの活用

PART1　データの収集と整理

問１　③→①→②

問２　（ア）④　（イ）①　（ウ）③　（エ）②

問３　（ア）○　（イ）○　（ウ）×　（エ）○

問４　エ

PART2　データの分析

問１　（ア）×　（イ）○　（ウ）×　（エ）×

問２　2019 年に比べて 2020 年の方が最大値，最小値，平均値いずれも高く，暑い日が続いたと考えられる。

問３　省略

問４　269，サッカー，女子

PART3　データの解釈１

問１　(1) 省略　(2) 省略

(3) 散布図が右肩上がり，相関係数が 0.9 以上であることから強い正の相関があるといえる。

問２　（ア）×　（イ）×　（ウ）○

PART4　データの解釈２

問１　(1) 省略　(2) 31577 円

問２　（ア）×　（イ）×　（ウ）×

問３　④→②→①→③

第 9 講　PART1　確認問題　解答

問 1

(1) ②

(2) ③

(3) ④

(4) ①

解説：

データの尺度には大きく, 間隔尺度, 比例尺度, 名義尺度, 順序尺度の 4 つがある。

問 2　①

解説：

オープンデータは，国や地方公共団体，事業者などが保有するデータを公開したものである。オープンデータには次の特徴がある。

・営利目的，非営利目的を問わず，二次利用可能

・コンピュータで容易に処理できるデータ形式

・無償で利用可能

問 3　④

解説：

さまざまな形をした，さまざまな性格を持った，さまざまな種類の膨大なデータのことをビッグデータという。その特徴は，データの量（Volume），データの種類（Variety），データの頻度・更新頻度（Velocity）の頭文字から「3V」と表現されることもある。ビッグデータを分析することでさまざまな意思決定に役立てることができる。

問 4

(1) ①

解説：

この処理をデータクレンジングという。

(2) ②

解説：

このような手法をデータマイニングという。